THE IMPLICATIONS FOR RESEARCH METHODOLOGY OF SOME BEHAVIORAL STUDIES IN PROGRAMED INSTRUCTION[1]

LEWIS D. EIGEN

The Center for Programed Instruction, Inc.

Since the time of E. L. Thorndike there has been no period in the history of education and psychology in which the two disciplines have been as inextricably connected as they are now with the advent of programed instruction. Although the attitude of the educator is usually "finally the psychologists have come to realize that they must deal with important problems in the real world," while the psychologists attitude is, "it's about time that these educators are going to let us show them how to do things sensibly"; there is, nonetheless, a great deal of professional contact and influence.

One of the more interesting areas of what might be described as "cross fertilization" is that of research methodology and procedures. Many of the research studies emanating from the programed instruction area of inquiry are unusual in that methodologically they do not resemble traditional educational studies nor are they parallels of typical studies in behavioral psychology. This paper shall be an attempt to analyze some of these studies and to demonstrate their implications for research methodology in education and psychology.

Student performance on examinations has been one of the standard criteria in educational experiments in the past. Learning has been measured in terms of the number of test items a student could correctly respond to after some experimental treatment. While test performance has also been a standard criterion for psychological experiments for many years, there has usually been an effort on the part of the experimenter to control for time and the number of trials taken to reach a criterion. Often the amount of time to reach a criterion, has been a standard measurement of learning (Howland, 1963). There are two main reasons why time is rarely, if ever, considered as a criterion in educational experiments. In the first place, the educational community, while very conscious of effectiveness, is rarely conscious of efficiency in its operation. Educators who might be concerned with efficiency are often labeled as advocates of quantity rather than quality. The second reason is an existing prejudice among educators that given any level of competence only a certain percentage of the human population could achieve that level. One of the objectives of the programer is always to prepare his materials so that virtually all students can learn. While this objective is rarely literally met, in some of the better programs it is approached very closely. Traditional teaching techniques have not in general been developed for the individual student, but programed instruction has changed that. This change makes the criterion of time a much more meaningful factor in evaluating learning. Thus, if it is true that all students of a given population will achieve complete mastery of a given task, then what is left to investigate but the time it takes to acquire this knowledge. While educational research has generally assumed acquisition of knowledge to be normally distributed, the ideal program of instruction would produce a single point on the distribution; all subjects will master the task to what-

[1]This paper was presented at a Conference on "Programed Instruction and Teaching Machines" in Berlin, Germany, July 9-15, 1963.

ever criterion is set. In an experiment comparing group with individual paces, Frye (1963) used as his criterion of learning the time it took students to learn to solve correctly two quadratic equations, one with a numerical coefficient and one with pro-numerical coefficients. He was attempting to compare the effects of a program when pacing was on an individual basis and on a group basis with both homogeneously and heterogeneously grouped subjects. Previous studies in this area had failed to find differences. One reason, perhaps, is that the level of learning was so uniformly high that all existing differences were subdued. Frye found that if heterogeneous grouping is used, individual pacing yields a more efficient result than does group pacing; if homogeneous grouping is used, there seems to be no difference in the efficiency of the two pacing techniques. Thus Frye demonstrated the differences that earlier experimenters had failed to detect. It should be noted that in his experiment, 31 of the 44 subjects who could not previously perform the required tasks were able to after their first run through the program. Eleven of the remaining 13 subjects of the program, and the two remaining subjects were successful after a second review.

Frye's study demonstrates a major influence and contribution of a traditionally psychological technique to educational research methodology by utilizing time as a criterion rather than just amount learned, which if used as the sole criterion runs the risk of subduing existing differences when the learning materials and/or techniques are of sufficiently high quality.

While the last experiment cited can be said to have given new importance to a type of criterion, some of the programed instruction studies have also minimized the importance of criteria which were traditional in educational experimentation. The attitude of the student has always been considered by educators to be of primary importance. The general view has been summarized by Fells and Trites as follows:

> There seem to be three necessary conditions for overt action to occur toward a particular goal, ... These are: (a) that the individual is capable of responding; (b) that the motive is dominant at the time; and (c) that the situation is favorable, that is, feasible, to the response. If all three conditions do not occur, action may be delayed, blocked, or diverted . . . .
>
> One of the most pervasive, . . . , effects of attitudes on behavior, both implicit (symbolic) and overt, involves their influences on and selective modifications of responses in particular situational settings (Harris, 1960, p. 109).

Because of this prevailing view student attitude has often been taken as one of, if not the sole, criteria of an experiment comparing educational techniques and/or materials. While many studies in programed instruction have reported student attitude in one way or another, the only two studies which correlate student attitude with student performance (time, as well as amount learned) find that the correlation is effectively 0 (Eigen, 1963; Feldhusen & Eigen, 1963).

Another interesting effect of the interaction of educator and psychologist in research with programed instruction occurs with respect to criteria for the success of teacher training techniques. In educational experiments, evaluation of teacher training techniques has taken the form of scores on written tests administered to teacher trainees; rating scales or checklists compiled by supervisors; interviews, visitations and conferences with teacher trainees; anecdotal records and diaries of the experiences in the classroom of teacher trainees; panels of juries, authorities and experts; case studies and the like. Rarely do we find teachers evaluated by means of how much their *students* learned or how efficiently they learned. While the thought of evaluating teachers or teacher training techniques by the amount that students

learn is completely onerous to many educators, this is precisely what is going on in a current experiment being conducted at Hunter College to determine if the experience of program writing, and studying about programed instruction will improve a prospective teacher's ability to teach. Here we see the influence of the behavioral psychologist not only in terms of what is being taught in the teacher training program (programed instruction), but also in terms of the criterion that is being used to determine whether or not the teachers will teach any better.

Criteria are by no means the only variables that have been affected. One of the traditional difficulties of educational research studies was the lack of control of the teacher. If a single teacher were used to administer two different experimental treatments it would be hard to keep unintentional bias out of the picture. If, on the other hand, different teachers were used, this presents another source of variation which must be accounted for. Statistical control has been the only possible solution, and this, of course, implies rather sizeable experiments in terms of the number of teachers and classes. A possible solution is the utilization of programs to present all experimental treatments. Gagne and Brown (1961) and Herrick (1962) have utilized this technique. In Gagne and Brown's case the attempt was to compare expository instruction with what they termed "guided discovery." Both sets of instructional materials were presented by means of programs; in Herrick's experiment the comparison was between a program with motivational problem setting questions appended at the begining of each unit, and a program without these questions. While the particular results of these studies are not very important so far as programed instruction is concerned, they serve to illustrate an influence that programed instruction is already having on educational experiments, in that, the teacher variables may be eliminated to a great extent by the use of properly designed programs.

Educators have been aware of the existence of individual differences for a long time. However, if one examines the behavior of the educational community with respect to their cognizance of individual differences, the findings will invariably be that they are essentially ignored in practice. Eigen and Komoski (1960) measured the time it took 77 students to proceed through a given program. The slowest student took approximately two and one-half times as long to finish the program as did the fastest student. The sample used in the experiment constituted a homogeneous population as grouped by the schools (IQ range of 105-135). Because for the past years education has operated essentially under a group pacing procedure, individual time differences such as those found in this study were difficult, if not impossible, to measure. Through the vehicle of programed instruction measurements of learning rate of academic materials became possible, and data such as this obtained by educational researchers and psychologists have virtually put an end to experiments dealing with concepts such as "a year's course," "a week's work," and the like. Thus, the educator is being forced to describe curriculum materials in behavioral rather than temporal terms which reflect a mean performance, as opposed to any individual's performance.

The "Montessori method" has often been advocated for the teaching of reading; so has the "Look-say method" and a myriad of others. The traditional means by which these two methods could be compared would be to have a sample of students taught by the "Montessori method" and another sample taught by the "Look-say method." Then the two groups of students would be compared with respect to

some criterion. This situation is typical of what might be described as the "pseudo-variable" or as Klaus has put it "gross variable" problem. The "Montessori method" of teaching reading is in actuality a conglomeration of many variables including pacing, response modality, reinforcement schedule, curriculum objectives, and the like. So is the "Look-say method" or any method, for that matter. Traditionally, the educator has been concerned with the very practical problem of which method of reading he should use to teach reading in his schools. The results of an experiment comparing two gross variables which are, in and of themselves, a compilation of many others may yield a decision for the educator, but it produces little understanding as to the processes and causes of the behavioral changes which take place. In essence, there are far too many variables which are uncontrolled. The educator reading the literature on programed instruction immediately runs across the "two schools of programing," the Skinner method and the Crowder method. Now, despite the fact that some researchers have attempted to compare these two gross variables to see which is the better method of programing the discipline of the behavioral psychologist has been felt with respect to this issue. In essence, the Skinnerian method of programing may be characterized by small steps, constructed responses, and a lack of branching; the Crowder method of programing, by large steps, multiple choice responses, and branching. Thus, there are at least three variables at issue and not one. There are a total of eight possible combinations of these three variables and not two. One might ask, "Why only contrast these two arrangements, what about the other six?" Essentially, there are two methods of controlling variation, experimental and statistical. An early experiment in the programed instruction literature which in my judgment is classic is one performed by Coulson and Silberman in 1959. In their study Coulson and Silberman investigated the three variables distinguishing the Skinnerian method of programing from the Crowderian method, response mode, step size and branching, but they controlled for the sources of variation by statistical means—a three way analysis of covariance. There was a significant superiority of small steps and a significant interaction of response modality with the branching variable. Thus, these researchers found significant effects which will aid in understanding the complex of variables which go to make up a program while, those studies comparing Skinner's method with Crowder's method (finding no significant differences) yield very little in the way of understanding and development in the field.

While Coulson and Silberman used a statistical method of control, other early researchers used experimental controls to avoid dealing with gross variables. An excellent example of this type of experiment is another study trying to investigate only the step size variable. Also in 1959, Evans, Glaser and Homme used 30, 40, 51 and 68 step sequences to attempt to teach the same material—conversions to number bases other than 10. In this experiment by dealing with a single variable and experimentally manipulating it the researchers obtained essentially the same results with respect to the variables they were dealing with as did Coulson and Silberman in their three-variable studies.

While there are still many pseudo or gross variable studies being conducted in education (and several in programed instruction), in general, educational researchers who may conduct a gross variable study in their first investigation with respect to programed instruction have generally modified their behavior and variables have

been controlled more effectively in the programed instruction literature than perhaps any other phase of educational research.

The behavior of the psychologist engaged in research in programed instruction has also been modified to a great extent. A majority of the leading psychologists doing research in programed instruction have come from the "Skinnerian" school. Evans (in press) has made the distinction between the usual investigating techniques of the Skinnerians, experimental analysis, as opposed to experimental design. The characteristics of experimental analysis include experimental laboratory control or elimination of variables as opposed to statistical control. Commonly, a single test subject is used, or very small numbers of subjects rather than large numbers. Also, results are usually presented graphically in the form of cumulative records as opposed to application of descriptive and inferential statistics and the use of probability values in summarizing results. Evans states that, "Although the impetus to the present flurry of research on teaching machines and programed learning was unquestionably provided by Skinner and other members of the experimental analysis school, practically all reported research . . . has been in the experimental design paradigm." There are several possible reasons for this state of affairs, not the least of which is the very nature of programs themselves. The typical graphical result of the cumulative record is in essence a function, mapping time elasped into the total number of correct responses. The frequency of response is restricted to the multiple occurrence of a single response in most instances. In a program of 100 frames it is conceivable that there may be 100 *different* responses each of which ostensibly is to be produced in the presence of a distinct, appropriate, discriminative stimulus. A cumulative record with a program under these circumstances would be relatively meaningless unless there were a meticulously careful analysis of each frame of the program to accompany it. Thus, the cumulative record in programed instruction research is of questionable value, and, in any event, is quite impractical to record.

As for the laboratory control, or elimination of variables, many psychologists who have been used to this luxury in their previous experiments have found that it is either impractical or impossible in many instances with programed instruction. Previous remarks in this paper have referred to statistical and experimental control of variables. The psychologist attempts to control, among other things, the behavioral repertoire of the organism prior to the experimental sessions. This, of course, is one of the great advantages of using infra-human animals as subjects for psychological experiments. The animals can be obtained from breeding farms where their entire lives have been controlled to a great extent prior to the time the experimenter receives them. While there are animal studies which have great relevance to the field of programed instruction, such as Terrace's (1963) study on discrimination learning without errors, programed instruction is essentially a method of controlling verbal behavior. This is true to a great extent even when programs attempt to teach psycho-motor skills. Thus, by and large, the experimenter in programed instruction, by necessity, must use human subjects. He is faced with a problem of great variability in the behavioral repertoires of these subjects. In human learning experiments of the past this difficulty was at least partially overcome by the judicious selection of materials on the part of the experimenter. Ideally, he would select nonsense syllables, numerical sequences, and the like, so that there was good reason

to believe that all subjects would have an equal lack of familiarity and facility with this material. Although this is desirable in programed instruction research, we must remember that since programed instruction is conceived of as a tool of great potential benefit to the education of the human race, there is a natural reluctance to use what might be described as "meaningless' or "irrelevant" materials in the experiments. The study by Frye referred to earlier is a typical example of the screening process which an experimenter may have to go through to acquire appropriate control prior to the experimental sessions. Frye's subjects had to be able to solve quadratic equations by means of factoring and at the same time must have lacked the ability to handle the quadratic formula. Another example of the pre-analysis of subjects that is often required is in a study of programing the concept of triangularity for very young children performed by Levin (Unpublished). Levin wanted to investigate properties of sequencing and their relationship to transfer. Prior to even developing the instructional materials, he had to conduct a rather elaborate study to determine what previous concepts of triangularity the experimental population had.

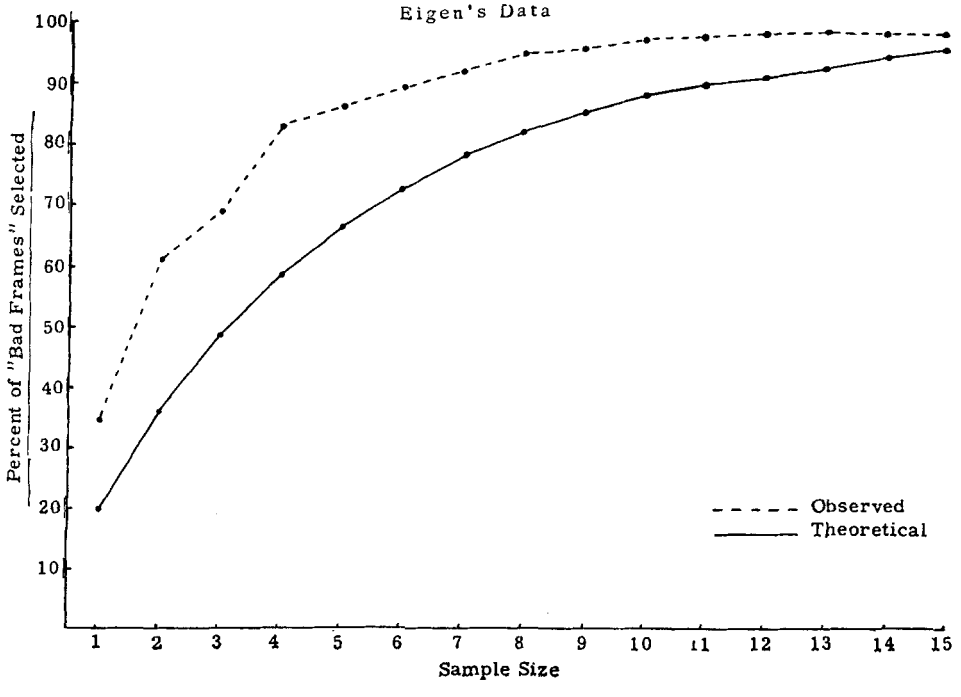
Since the verbal behavior which is the end product of academic education or industrial training is usually of such complex magnitude, it is not only difficult to control or eliminate many experimental variables, but it may be indeed unwise to do so even if possible, since there may be highly significant interactions among experimental variables. The significant interaction obtained by Coulson and Silberman between response modality and branching is one example. If the experimenters in this case had eliminated additional variables and/or controlled for them experimentally, the interaction might not have been evident. Another example of this phenomenon can be given with respect to the overt-covert issue. Many studies have been performed failing to find significant differences between overt and covert responses. In most of these studies all other experimental variables were eliminated or controlled experimentally. In this particular case it is an easy task simply to take a program and instruct the subjects to respond overtly, by writing their responses, or covertly, by thinking about them. In a recent experiment in which not only was the overt-covert variable investigated, but also, the relative merits of incidental and elevant responses, Eigen and Margulies (1963) examined, each of these combinations of treatments for material of three different information levels (information level is a mathematical concept to measure how far the response is from the repertoire of the subject). As is evident for incidental responses there is a difference between levels of information. However, when the responses in the program are relevant (as is theoretically sound) then the overt response is significantly superior to the covert response beyond the .01 level only for the intermediate and high information levels but not for the low information level. This, in effect, suggests that where subjects are relatively familiar with the responses, it is not important whether one requires overt or covert responses; but when these responses are unfamiliar, the overt response mode is superior. Here we have the result of a three-way interaction; it has definite implications for how programs should be written and used. But if the variables had been experimentally controlled or eliminated, there is doubt as to whether any light would have been shed on the subject.

Another area in which the behavior of the experimental psychologist has been changed with respect to the experimental methodology is in the numbers of subjects used. Since the psychologist has not relied on cumulative records or other simple

graphical methods of presentation of results and has instead utilized the experimental design approach necessitating statistical inference, he is put at a gross disadvantage when an experiment performed with very small numbers of subjects fails to achieve significant differences. He has no way of knowing whether or not his findings are simply due to the small sample size or due to a real lack of difference among the variables under consideration. An example of this problem can be seen in one of the early overt-covert studies (Evans et al., 1960). In this experiment, five subjects were instructed to write their answers, and five subjects, to think their answers. The existing differences failed to reach statistical significance, and, as a result, the experiment was not particularly successful in terms of clarifying the overt-covert issue. It should be pointed out that even though an inferential statistical technique was utilized, small numbers of subjects do not cause any problem if significant differences are obtained. It is only where they are not obtained that the reader of the experiment is in a dilemma which might well have been eliminated had larger numbers of subjects been used.

Even in the case where the experimental analysis technique might be used (this is usually the technique adopted when a program is being developed) a recent experiment suggests that the use of a single subject or very small numbers of subjects can produce undesirable consequences. In this study (Eigen, Unpublished) the experimenter analyzed the responses of 33 subjects to each frame of a program. He then defined a master list of "bad frames" (frames indicative of a fault of the program) by including any frame on which 20% or more of the population had made an error. Then, random samples of size 1 through 15 were drawn by means of a com-

FIGURE 1



puter, and the results of the samples were compared with the master list to determine what percentage of "bad frames" would have been detected by the experimenter had he only used a sample size of each particular value; and if one or more of the subjects had made an error, the frame would have been classified as a "bad frame." Figure 1 summarizes this analysis. Note that if only one subject were used (a typical experimental analysis procedure) then 64% of the "bad frames" would have gone undetected, while if three subjects were used approximately 30% would have been undetected. Thus, due to the great variability of individual conditioning histories prior to the investigation, reliance on extremely small samples is a questionable practice. On the other hand, six subjects produced a mean of 90% and eight, a mean of over 95%. Thus, a law of diminishing returns is in effect, and the necessity of using large samples, 20 or more (the usual experimental design technique), is also of questionable value.

As we have seen with programed instruction the educational and the psychological researcher have come into intimate contact. The practices of each group have been modified by those of the other, as well as the unique properties and problems of programed instruction. This phenomenon is extremely desirable, in that, it enables the programed instruction movement to benefit from the best of both worlds.

References

Coulson, J. E., & Silberman, H. F. Results of an initial experiment in automated teaching. In R. Glaser & A. Lumsdaine (Eds.), *Teaching machines and programed learning.* Washington, D. C.: Nat. Educ. Ass., 1960. Pp. 452-468.

Eigen, L. D. An empirical approach to the determination of sample size used in the development of a program. Unpublished study done at The Center for Programed Instruction, New York.

Eigen, L. D. High school student reactions to programed instruction. *Phi Delta Kappan,* 1963, *54,* 282-285.

Eigen, L. D., & Komoski, K. *Research summary #1.* New York: Center for Programed Instruction, 1960.

Eigen, L. D., & Margulies, S. Response characteristics as a function of information level. *J. programed Instruction,* 1963, *2* (in press).

Evans, J. L. Programming in mathematics and logic. In R. Glaser & A. Lumsdaine (Eds.), *Teaching machines and programed learning,* Vol. 2. Washington, D. C.: Nat. Educ. Ass., in press.

Evans, J. L., Glaser, R., & Homme, L. E. A preliminary investigation of variations in the properties of verbal learning sequences of the "Teaching Machine" type. In R. Glaser & A. Lumsdaine, (Eds.), *Teaching machines and programed learning.* Washington, D. C.: Nat. Educ. Ass., 1960. Pp. 446-451.

Feldhusen, J., & Eigen, L. D. Interrelationships among attitude, achievement, reading, intelligence, and transfer variables in programed instruction. Paper read at Midwest. Psychol. Ass., Chicago, May, 1963.

Frye, C. H. *Group versus individual pacing in programed instruction.* Portland, Oregon: Oregon State System of Higher Education, Teaching Research Project, 1963.

Gagne, R. M., & Brown, L. T. Some factors in the programming of conceptual learning. *J. exp. Psychol.,* 1961, *62,* 313-321.

Harris, C. W. (Ed.) *Encyclopedia of educational research.* New York: Macmillan, 1960.

Herrick, M. C. *The effect of problem-setting questions on rate and amount of learning in programming teaching machines.* Bloomington, Indiana: Indiana Univer. Audiovisual Center, 1962.

Hovland, C. I. Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1963. Pp. 613-689.

Levin, J. R. Children's concept of triangularity. Unpublished paper written at Brown University, Providence, Rhode Island.

Terrace, H. S. Discrimination learning with and without errors. *J. exp. anal. Behav.,* 1963, *6* (1), 1-27.